Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.896
DOI: https://doi.org/10.24867/05BE41Nikolic

SISTEM ZA UPRAVLJANJE ŽALBAMA KORISNIKA U DOMENU FINANSIJA

CUSTOMER COMPLAINTS MANAGEMENT SYSTEM IN THE FIELD OF FINANCE

Milica Nikolić, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *U radu je predstavljen sistem za upravljanje žalbama korisnika u domenu finansija. Sistem je zasnovan nad velikim brojem realnih žalbi sa ciljem da kompaniji pruži mogućnost automatskog upravljanja žalbama, tako što će u realnom vremenu vršiti predikcije koje rešene žalbe ne zadovoljavaju korisnike koji su ih uložili i koji je očekivani broj žalbi za naredni mesec. Krajnji cilj ovog sistema jeste smanjenje broja nezadovoljnih korisnika date kompanije. Zbog rada sa velikom količinom podataka, ceo sistem odvija se u distribuiranom okruženju.*

Ključne reči: *određivanje sentimenta, klasifikacija žalbe, rad sa velikom količinom podataka, distribuirano okruženje*

Abstract – *This paper presents the customer complaints management system in the field of finance. The system is based on large number of complaints, with the primary goal of providing the ability of automatic management to the company, by predicting complaints which are likely to be disputed and the number of expected complaints for the following month. The final goal of the system is to reduce the number of unsatisfied customers of the specific company. Since this is a big data problem, the whole system is running in a distributed environment.*

Keywords: *sentiment mining, complaints classification, big data, distributed environment*

1. UVOD

Ulaskom u novu digitalnu eru, započet je proces jednostavnijeg prikupljanja, elektronske razmene i skladištenja informacija.

U cilju zaštite svojih građana, američka vlada osnovala je zavod za finansijsku zaštitu potrošača, čije jedno od odeljenja predstavlja odsek za žalbe korisnika [1]. Svaki građanin, preko veb sajta ovog zavoda može da uloži žalbu na kompaniju sa kojom sarađuje, nakon čega kompanija ima određeni rok da razreši žalbu i da pruži odgovor, koji potom može biti prihvaćen ili osporen od strane korisnika.

Na ovaj način, građani mogu da saznaju koje kompanije su otvorene ka rešavanju problema svojih korisnika, a koje imaju nezadovoljne korisnike. Sa druge strane, kompanije imaju uvid u poslovanje svojih konkurenata.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, docent.

Cilj ovog sistema jeste automatsko upravljanje žalbama korisnika na način da pruži krajnjem korisniku – kompaniji uvid u postojeće žalbe i da bude sposoban da izvrši predikcije intenziteta žalbi za naredni mesec, kao i da uputi kompaniju na one žalbe koje imaju veću mogućnost da budu osporene.

Shodno cilju, sistem je podeljen na dva velika, nezavisna podsistema:

- podsistem za predickiju intenziteta žalbi u narednom mesecu
- podsistem za predikciju odgovora korisnika na razrešenu žalbu – da li će korisnik da prihvati odgovor kompanije ili će da ga ospori.

Ovi podsistemi naslanjaju se na podsistem za prikupljanje i skladištenje podataka.

Predikcija intenziteta žalbi, odnosno broja žalbi u narednom mesecu, zasniva se na upotrebi modela linearne regresije (engl. *Linear Regression*). Kao jedini atribut koristi se broj žalbi iz prethodnog meseca, kako bi se predvideo njihov broj u narednom. Iako predikcije ovog modela nisu naročito precizne, sistem je u stanju da predvidi smer kretanja intenziteta žalbi i da približan opseg broja žalbi u narednom mesecu u veoma kratkom vremenskom periodu, što je od izuzetnog značaja za kompaniju, te, u ovu svrhu, drugi modeli nisu isprobavani.

Za predviđanje osporavanja odgovora kompanije od strane korisnika, isproban je *kernel-based* algoritam - *Support Vector Machine (SVM),* kao i nekoliko algoritama baziranih na principu stabala odlučivanja. Svi modeli isprobani su nad istim ulaznim skupom podataka koji se sastoji iz sledećih atributa – usluga i problem na koju korisnik ulaže žalbu, odgovor kompanije, atribut koji govori da li kompanija dozvoljava javno objavljivanje odgovora, sentiment teksta žalbe korisnika, atribut koji govori da li korisnik dozvoljava javno objavljivanje tekstualnog opisa žalbe, lokacija sa koje je uložena žalba, način na koji je žalba podneta (usmeno, putem telefona, imejla, itd.), atribut koji govori da li je kompanija pružila odgovor na žalbu u dogovorenom vremenskom roku i atribut koji označava da li je korisnik osporio odgovor kompanije ili ga je prihvatio.

Za implementaciju ovog rešenja, prednost su dobili modeli zasnovani na stablima odlučivanja zbog činjenice da imaju mogućnost da prikažu udeo atributa u konačnoj predikciji, što je od izuzetne važnosti za krajnjeg korisnika (kompaniju).

Najbolje performanse pokazao je *XGBoost (eXtreme Gradient Boosting)* algoritam, što ga je odredilo za konačno rešenje ovog sistema.

2. METODOLOGIJA

Programsko rešenje implementirano je u *Python*-u, pri čemu se ceo proces odvija u distribuiranom okruženju. Za kontrolu rada distribuiranog sistema i za skladištenje podataka, upotrebljena je sistemska platforma *Hadoop* [2], dok se obrada podataka, treniranje modela i dobijanje predikcija vrše pomoću veoma popularnog alata za distribuirano programiranje – *Spark*-a [3]. Logika sistema enkapsulirana je u mikroservis koji putem *REST API*-ja komunicira sa *Angular* [5] veb aplikacijom, preko koje korisnik ostvaruje vezu sa sistemom, izvršava zadatke i dobija rezultate.

Sistem se sastoji iz tri celine. Podsistem za prikupljanje i skladištenje podataka osnova je za rad preostala dva podsistema, koji služe za rešavanje dva glavna problema ovog sistema i čine ih:

- podsistem za predickiju intenziteta žalbi u narednom mesecu
- podsistem za predikciju odgovora korisnika na razrešenu žalbu

2.1. Podsistem za prikupljanje i skladištenje podataka

Podsistem za prikupljanje i skladištenje podataka zadužen je za prikupljanje i skladištenje velike količine podataka, pri čemu se ažuriranje podataka obavlja na dnevnom nivou. Komunikacija i preuzimanje podataka sa zvaničnog sajta [1] vrši se upotrebom *Socrata API*-ja [6] i zasnovana je na slanju *SoQL (Socrata Query Language)* upita, jezika koji se u potpunosti oslanja i predstavlja dopunu *SQL*-a (*Structured Query Language*). S obzirom na to da je broj zapisa višemilionski, podaci se skladište na distribuiranom fajl sistemu - *Hadoop Distributed File System (HDFS).*

2.2. Podsistem za predikciju intenziteta žalbi u narednom mesecu

Podsistem za predikciju intenziteta, odnosno broja žalbi u narednom mesecu zasnovan je na modelu linearne regresije [7]. Model je izuzetno jednostavan, bez značajnijih meta-parametara, te ne zahteva dodatne pretrage u svrhu traženja najbolje kombinacije njihovih vrednosti.

Sve prikupljene žalbe grupišu se po mesecu i računa se njihov broj, a potom se taj broj i broj žalbi iz narednog meseca (koji predstavlja ciljno obeležje) šalju kao ulaz modelu. Na ovaj način, model se obučava da predvidi broj žalbi u narednom mesecu, ako mu je poznat samo broj žalbi iz trenutnog meseca.

2.3. Podsistem za predikciju odgovora korisnika na razrešenu žalbu

Podsistem za predikciju odgovora korisnika na razrešenu žalbu predstavlja složeniji podsistem i može se posmatrati kao sistem za sebe. Sastoji se iz dva modula:

- podsistem za analizu i obradu podataka
- podsistem za pronalaženje meta-parametara i treniranje modela

2.3.1. Podsistem za analizu i obradu podataka

Podsistem za analizu i obradu podataka kao ulaz dobija višemilionski broj neobrađenih žalbi, vrši njihovo procesiranje, a potom tako obrađene žalbe šalje na izlaz, odnosno ulaz podsistema za pronalaženje meta-parametara i treniranje modela. Podsistem za analizu i obradu podataka rešava dva velika problema:

- selekcija obeležja
- eksplorativna analiza i priprema ulaznih podataka za treniranje modela.

Model za prikupljanje i skladištenje žalbi skuplja i skladišti žalbe u originalnom formatu, tačno onako kako su objavljene na zvaničnom sajtu [1]. Ovakav skup opisan je pomoću 18 atributa, shodno čemu količina vremena potrebna za njegovu obradu i treniranje modela i procesna moć računara za rad sa ovim podacima rastu. Nad skupovima podataka sa velikim brojem atributa često se vrši selekcija obeležja, pri čemu se znatno smanjuje vreme potrebno za izvršavanje datog problema, što dovodi do toga da je sistem u stanju da pruži odgovor u realnom vremenu.

Selekcija obeležja implementiranog sistema svodi se na dobro poznavanje domena i posmatranje korelacije između različitih atributa. Najpre se iz osnovnog skupa podataka eliminišu atributi poput rednog broja žalbe, oznake koja je dodeljena žalbi i slično, s obzirom na to da ovakvi atributi nemaju uticaj na buduću predikciju. Potom se posmatraju rezultati korelacije između različitih atributa, na koji način dolazi do izbacivanja atributa. Na ovaj način izbačeni su zip kod (koji je u korelaciji sa atributom koji predstavlja lokaciju sa koje je podneta žalba), datum pristizanja i datum prosleđivanja žalbe kompaniji (koji su u korelaciji sa atributom koji govori da li je odgovor na žalbu dat u okviru dogovorenog vremenskog perioda) i još neki slični atributi koji se mogu opisati upotrebom nekih od preostalih atributa skupa podataka. Konačan skup atributa predstavljen je u poglavlju 2.4.

Eksplorativna analiza obuhvata detaljnu analizu, često praćenu grafičkim prikazima, svih atributa ulaznog skupa i svih njihovih mogućih vrednosti. Primenom eksplorativne analize, došlo se do važnih statističkih otkrića poput tih da su svi atributi ulaznog, redukovanog skupa kategorički, da se skup sastoji iz velikog procenta nedostajućih vrednosti i da je skup vrednosti ciljne promenljive znatno neizbalansiran.

Binarni kategorički atributi, koji često daju odgovore na da/ne pitanja poput toga da li je žalba osporena, da li je kompanija pružila odgovor u dogovorenom roku, da li je korisnik pristao da se tekst žalbe objavi javno – pretvoreni su u numeričke, prostom dodelom broja 1 kao odgovora "da", odnosno 0 kao odgovora "ne", dok su preostali kategorički atributi, čiji broj mogućih vrednosti kategorija prelazi i preko stotinu, pretvoreni u numeričke attribute upotrebom *Label Encoding*-a [8].

Problem nedostajućih vrednosti rešen je popunjavanjem vrednosti atributa pomoću vrednosti nekih drugih atributa, na primer, nedostajuće vrednosti lokacija dobijene su upotrebom vrednosti zip kod atributa. Instance sa velikim brojem nedostajućih vrednosti atributa eliminisane su iz skupa podataka. Jedan od atributa sa najvećim brojem nedostajućih vrednosti jeste upravo sam tekst žalbe, koji bi trebalo da predstavlja glavni atribut prilikom predikcije

da li će korisnik prihvatiti odgovor kompanije ili će ga osporiti. Ovaj atribut, predstavljen u tekstualnoj formi, takođe je nepogodan za treniranje izabranog modela, koji ne podržava rad sa tekstualnim podacima, te je atribut zamenjen novim atributom koji predstavlja numeričke vrednosti sentimenta žalbe, dobijene primenom *VADER (Valence Aware Dictionary for sEntiment Reasoning)* [9] algoritma nad datim tekstom. Za vrednosti sentimenta žalbi čiji tekstualni opis nije dat, upotrebljena je vrednost 0, koja označava da tekst ove žalbe nije ni pozitivan ni negativan, s obzirom na to da bi izbacivanje svih instanci sa nedostajućim vrednostima ovog atributa znatno smanjilo skup podataka za treniranje.

Eksplorativna analiza pokazala je i veliki disbalans kada je u pitanju broj osporenih i neosporenih žalbi, tačnije broj neosporenih žalbi daleko je veći od broja osporenih. Za rešavanje ovog problema, primenjena je tehnika *SMOTE (Synthetic Minority Over-sampling Technique)* [10], koja predstavlja jednu od tehnika *oversampling*-a.

2.3.2 Podsistem za pronalaženje meta-parametara i treniranje modela

U svrhu predikcije korisnikovog odgovora na razrešenje žalbe od strane kompanije, iskorišćen je *XGBoost (eXtreme Gradient Boosting)* [11] algoritam. Performanse ovog modela znatno se mogu poboljšati pronalaženjem odgovarajuće kombinacije vrednosti meta-parametara. Za pronalaženje meta-parametara modela koji će rezultirati najboljim performansama, upotrebljena je kombinacija *RandomizedSearchCV* [12] i *GridSearchCV* [13] algoritma.

Postupak traženja ovakvih meta-parametara započinje upotrebom *RandomizedSearchCV* algoritma koji kao početnu mrežu parametara prima parametre modela sa vrednostima u predefinisanim opsezima.

RandomizedSearchCV vrši nasumičnu pretragu parametara, gde svaka kombinacija predstavlja uzorak distribucije mogućih vrednosti parametara. Po izvršavanju ovog algoritma, dobija se *XGBoost* model sa "najboljim" performansama kao i vrednosti meta-parametara koje su rezultirale ovim modelom. Ovako dobijene vrednosti meta-parametara često ne predstavljaju najbolje već približno najbolje vrednosti, dobijene uz znatno mali utrošak vremena. Potom se, za dobijanje dodatne tačnosti, koristi *GridSearchCV* algoritam, koji se zasniva na isprobavanju svih mogućih vrednosti parametara.

Kao ulazna mreža vrednosti parametara za pretragu koriste se vrednosti dobijene upotrebom *RandomizedSearchCV* algoritma kao i vrednosti iz njihove neposredne okoline. Ovako istreniran model pokazuje najbolje performanse i prihvata se kao konačno rešenje. Predikcije se izvršavaju u realnom vremenu, dok treniranje modela zahteva veću količinu vremena i odvija se u odloženom režimu. Budući da se podaci svakodnevno ažuriraju, model se ponovo trenira na mesečnom nivou.

2.4. Skup ulaznih podataka

Od 18 ulaznih atributa originalnog skupa podataka, 14 atributa ima kategoričke vrednosti, jedan predstavlja tekstualni opis, dok su dva tipa datuma i samo jedan atribut je numerički. Većina kategoričkih atributa velikog je kardinaliteta – broj mogućih kategorija prelazi preko sto.

Kao ulazni skup podsistema za predikciju intenziteta žalbi u narednom mesecu, uzima se samo datum objavljivanja žalbe na sajtu, tj. datum pristizanja žalbe.

Za pripremu skupa za predikciju odgovora korisnika na razrešenje žalbe, prvo se vrši obrada i analiza podataka koju čine selekcija obeležja i eksplorativna analiza, opisane u poglavlju 2.3.1, nakon čega se skup podataka sastoji iz atributa čiji su opisi dati u tabeli 2.1.

Tabela 2.1: Ulazni skup atributa *XGBoost* modela

ATRIBUT	OPIS ATRIBUTA
product	usluga na koju je žalba upućena
sub_product	pod-kategorija usluge na koju je žalba upućena
issue	problem koji korisnik ima
sub_issue	pod-kategorija problema koji korisnik ima
vader_sentiment	numerička vrednost sentimenta žalbe
company_public_response	da li kompanija dozvoljava javno objavljivanje odgovora
state	lokacija sa koje je žalba upućena
consumer_consent_provided	da li korisnik dozvoljava javno objavljivanje tekstualnog opisa žalbe
submitted_via	način na koji je žalba upućena zavodu
company_response	odgovor kompanije na žalbu
timely	da li je kompanija pružila odgovor u dogovorenom roku
consumer_disputed	da li je korisnik osporio odgovor kompanije

Ovaj skup atributa čini ulazni skup u *XGBoost* model, pri čemu atribut *consumer_disputed* predstavlja ciljnu promenljivu.

3. EVALUACIJA

Verifikacija celokupnog sistema u suštini se svodi na verifikaciju podsistema za predikciju intenziteta žalbi za naredni mesec i podsistema za predikciju korisnikovog odgovora na razrešenu žalbu od strane kompanije. Tačnije, verifikacija sistema podrazumeva verifikaciju modela linearne regresije i *XGBoost* modela.

Model linearne regresije evaluira se podelom ulaznog skupa na trening i test skup u odnosu 85% na prema 15%, upotrebom vrednosti R^2 mere, koja iznosi 0.64. Iako ovaj model ne daje izuzetno precizna rešenja, empirijskim putem je ustanovljeno da je model u stanju da predvidi smer kretanja intenziteta žalbi u narednom mesecu, tj., da predvidi da li će njihov broj naglo opasti ili porasti, kao i da pruži okviran opseg vrednosti za očekivani broj žalbi u narednom mesecu. Sa druge strane, model je izuzetno jednostavan, bez dodatnih parametara za podešavanje i kao takav ne zahteva veću količinu vremena potrebnu za njegovo treniranje i predikciju rezultata.

Za treniranje *XGBoost* modela koji će pružiti najbolje performanse, vrše se dodatne pretrage, opisane u poglavlju 2.3.2, koje pronalaze takvu kombinaciju vrednosti meta-parametara da istrenirani model rezultira najboljim performansama. Proces pronalaženja ovakvog

modela, sastoji se iz višestrukog treniranja *XGBoost* modela, svaki put sa različitim kombinacijama vrednosti meta-parametara, pri čemu se svaki model evaluira metodom unakrsne evaluacije tako što se ulazni skup deli na *k* segmenata podeljenih na trening i test skup, a potom se kao konačna tačnost ovog modela uzima prosečna tačnost svih ovih segmenata. Ovakav način evaluacije izabran je zbog činjenice da *RandomizedSearchCV* i *GridSearchCV* algoritam pružaju direktnu podršku za unakrsno evaluiranje, primajući ovu metodu kao jedan od svojih parametara. Kao mere evaluacije koriste se tačnost i odziv. Prvenstveno se traži model sa najvećim odzivom, smatrajući da nekoliko lažno klasifikovanih žalbi ne predstavlja veliku grešku sve dok su sve žalbe koje treba da budu klasifikovane kao osporene - klasifikovane tačno, ali ipak vodeći računa da tačnost modela ne spadne ispod empirijski, unapred određene granice od 70%[1]. S obzirom na to da su vrednosti meta-parametara svaki put nasumično odabrane, tačnost ovog modela je promenljiva ali se kreće u neznatno malom opsegu, pri čemu odziv uvek prelazi 85%, dok je tačnost iznad 75%.

Ovaj model pruža mogućnost prikaza udela svakog od pojedinačnih atributa u predikciji. Grafik udela atributa pokazuje da način prilaganja žalbe kao i činjenica da li je kompanija dozvolila javno objavljivanje svog odgovora imaju veliki uticaj na predikciju. Takođe, neočekivano, lokacija sa koje je žalba podneta ima veliku ulogu prilikom predikcije, dok razlozi žalbe i činjenica da li je kompanija odgovorila u zadatom vremenskom roku ne utiču puno na predikciju da li će korisnik osporiti odgovor na žalbu.

Modeli opisani u ovom radu specifično se treniraju nad skupom podataka vezanim za isključivo jednu kompaniju, te i da postoji sistem sličan ovom, verovatno je razvijen u okviru same kompanije i u postojećoj literature nije pronađen nijedan rad sa istim ciljem predviđanja, vezan za implementaciju ili rezultate ovakvog sistema, te je teško porediti rezultate sa postojećom literaturom.

4. ZAKLJUČAK

U radu je predstavljen model sistema za upravljanje žalbama korisnika u domenu finansija. Logika celog sistema enkapsulirana je u mikroservis koji putem *REST API*-ja komunicira sa veb aplikacijom pomoću koje krajnji korisnik – kompanija, izvršava određene zadatke i dobija odgovore od sistema. Aplikacija je zasnovana na realnim, javno dostupnim podacima i služi da omogući kompaniji uvid u sve do tada njoj upućene žalbe i da izvrši predikcije o intenzitetu žalbi u narednom mesecu i odgovoru korisnika na razrešenu žalbu, u realnom vremenu. U svrhu predikcije intenziteta žalbi iskorišćen je model linearne regresije, koji je bez ikakvog podešavanja parametara rezultirao vrednošću *R2* mere od 0.64. Za predikciju odgovora korisnika na razrešenje žalbe, upotrebljen je *XGBoost* model, čije vrednosti parametara su određene kombinacijom *RandomizedSearchCV* i *GridSearchCV* algoritma. Ovaj model evaluiran je metodom unakrsne evaluacije, pri čemu se kao metrike evaluacije uzimaju odziv i tačnost. Vrednost mere odziva prelazi 85%, dok je tačnost iznad 75%. Dodatna prednost *XGBoost* modela jeste ta da je u stanju da prikaže udeo svakog pojedinačnog atributa u predikciji, što je, pored mogućnosti oba modela da pružaju predikcije u realnom vremenu, od izuzetnog značaja za krajnjeg korisnika (kompaniju).

Problem predikcije intenziteta žalbi može se posmatrati kao problem analize vremenskih serija, te bi upotreba modela poput *ARIMA* modela, mogla pružiti preciznije predikcije i na taj način poboljšati ovaj deo sistema. Još jedan od mogućih pravaca unapređenja sistema jeste drugačiji odabir klasifikatora prilikom predikcije korisnikovog odgovora na razrešenu žalbu. Nedavno objavljeni - *BERT* model [15] pokazuje *state-of-the-art* rezultate na *NLP (Natural Language Processing)* baziranim problemima.

[1] Granica je određena zahtevom kompanije koja koristi razvijeni sistem.

5. LITERATURA

[1]https://www.consumerfinance.gov/data-research/consumer-complaints/ [pristupljeno 16.7.2019.]
[2]https://hadoop.apache.org/ [pristupljeno 15.7.2019.]
[3]https://spark.apache.org/ [pristupljeno 15.7.2019.]
[4]https://restfulapi.net/ [pristupljeno 16.7.2019.]
[5]https://angular.io/ [pristupljeno 16.7.2019.]
[6]https://dev.socrata.com/foundry/data.consumerfinance.gov/s6ew-h6mp [pristupljeno 16.7.2019.]
[7]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html [pristupljeno 16.7.2019.]
[8]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html [pristupljeno 15.7.2019.]
[9] Clayton J. Hutto / Eric Gilbert, *Vader: A parsimonious rule-based model for sentiment analysis of social media text*, In: Eighth international AAAI conference on weblogs and social media, 2014.
[10]https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html [pristupljeno 16.7.2019.]
[11]https://xgboost.readthedocs.io/en/latest/ [pristupljeno 16.7.2019.]
[12]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html [pristupljeno 16.7.2019.]
[13]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [pristupljeno 16.7.2019.]
[14] Jacob Devlin et al, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, 2018.

Kratka biografija:

Milica Nikolić rođena je u Novom Sadu 1994. godine. Osnovne akademske studije iz oblasti *Računarstvo i automatika*, završila je na Fakultetu tehničkih nauka, 2017. godine. kada upisuje i master akademske studije iz iste oblasti. Master rad odbranila je 2019 godine.

kontakt: milka94ns@gmail.com